Abstract

This is a proposal to examine the existence of, and to assess the possible impact of, the effects of the clustering of items around reading passages on NAEP outcomes.

It is well known that the NAEP employs item response theory (IRT) in its analytic procedure to process item responses. One of the basic premises of IRT is the local independence among item responses given student proficiency. Such an assumption is hard to justify when items are clustered around a reading passage, as in the Reading Assessment, or around a hands-on science project, as in the Science Assessment. Information derived from a cluster is typically less than that derived from locally independent items. In the extreme case, information derived from asking several slightly differently phrased questions is almost equivalent to that from asking a single question. This could lead to an overestimation of the precision in student proficiency. We have collected empirical evidence from an earlier study that local dependency (deviation from the local independence assumption) exists in the 1994 Long Term Trend Reading Assessment. With the growing use of item clusters in NAEP, it has become pressing to develop analytic procedures that enable one to assess the impact of passage effect, and to use these procedures to assess the actual impact on NAEP reported proficiency scores.

We propose four studies. In the first study, we propose a rigorous statistical procedure to identify and measure local dependency in NAEP data. We have already performed extensive analysis on the Long Term Trend Reading Assessment in this study. In the other studies, we propose to compare NAEP outcomes under the locally dependent and locally independent models. Local dependency is captured via the use of a statistical model. Because of the introduction of this new component of dependency structure into the already sophisticated NAEP methodology, the resulting analytic procedure becomes computationally difficult. We propose to solve the problem, within our practical limitation of resources, first by approximating the NAEP methodology using a Bayesian approach, and secondly, by limiting the scope of our study. The Bayesian framework encompasses both the locally independent and locally dependent models, and it enables

us to use powerful computational tools to perform the necessary inference. With the aid of a simulated response data set modeled after the Long Term Trend Reading Assessment, we demonstrate how these Bayesian tools are deployed to solve the locally independent and locally dependent models.

Contents

Assessing the Psychometric Effects of Item Clustering Around Passages in the National Assessment of Educational Progress

Eddie Ip
Steven Scott

1. Objective

The objective of the proposed project is to investigate the existence of passage effect and its possible impact on the precision of reported NAEP results. That is, does the clustering of items around a passage or task have a psychometric effect on reported proficiency scores, and, if it does, in what way and by how much. In the process, we propose to develop, under a rigorous statistical framework, analytic procedures directed to this purpose.

2. Statement of the Problem

Item clusters are ubiquitous in NAEP assessments. An item cluster is defined as a collection of two or more items that share a common reading passage. The reading passage could be a literary passage from the NAEP Reading Assessment, or a description of a scientific experiment in the NAEP Science Assessment. The use of item clusters in NAEP assessments seems to be increasing over the years. The Science Assessment, for example, following guidelines suggested by the NAEP Governing Board (1996), recently began to pilot test item blocks – large item clusters that are centered on one task or hands-on experiment. A student is asked to perform a simple experiment, using a standard experiment kit. Then the student answers several questions related to that experience. The rise in popularity of item clusters is understandable and is a welcome trend. Multiple choice tests that contain only standalone items can hardly be used to test a student's ability to organize, connect and integrate information. Item clusters, on the other hand, may be more flexible in testing higher cognitive skills such as integration of concepts and recognizing connections between various pieces of information. For example, items from a reading passage of the ecology of a pond may test how a student's understanding of science can lead to his understanding of its societal impact.

While it is understandable that clustered items are well suited for testing higher cognitive skills, the possible correlation between responses within an item cluster given a particular level of student proficiency creates a psychometric problem and also possible problems on the precision of reported NAEP results. It is well known that NAEP uses item response theory (IRT) to scale items (Beaton & Zwick, 1992). One basic premise in IRT is the local independence assumption that the responses are conditionally independent given ability or proficiency (Lord and Novick, 1968). Loosely speaking, this means that the correlation between a student's item responses can be completely explained by the proficiency of the student and no other factors.

The assumption of local dependency (LI) seems hardly defensible when a NAEP assessment consists of one or more item clusters. Consider the following hypothetical example. A Reading Assessment consists of ten item clusters – five of them are reading passages concerning "boy activities" such as baseball and automobiles, and the other five concern "girl activities" such as home economics and fashion design. A boy and a girl who both correctly respond to, say, 70% of the items may be judged to have approximately the same reading proficiency. However, a closer look at their response patterns may reveal that the boy tends to correctly answer the items in clusters regarding "boy activities," and the girls correctly answer "girl activities" items. Specifically, we can say that there exists a student-cluster interaction effect. In psychometric terms, given reading proficiency, item responses do not satisfy the condition of LI.

Mathematically, the LI assumption can be stated as :

$$P(Y=y|q)=P(Y_1=y_1|q)\cdots P(Y_J=y_J|q), \qquad (1)$$

where Y_j is the response to item j, q is the unobserved student proficiency, and $Y=(Y_1,\cdots,Y_J)$. Deviation from the local independency condition specified by equation (1) will be referred to as local dependency (LD).

How does LD affect a test consisting of clustered items? Wainer(1995) points out that items in a cluster are to some extent redundant. He states:

> To the extent that the response to the second item in a pair depends on (and, therefore, can be predicted from) the response to the first item, beyond prediction from the underlying proficiency being measured, the

> second item provides less information than would a completely (locally) independent item. (p. 158)

Consider the following extreme case: if all the items in a cluster of $J>1$ related items are almost perfectly correlated (for example, when the same question is phrased slightly differently several times), then the information in this cluster effectively reduces to the information from a single item. The rest of the items are redundant. Statistically, the reduction of information due to LD in a test implies that the standard error of the ability estimate as measured by the test under the LI assumption is understated. In other words, the precision of the reported student score is overstated. Under maximum likelihood (ML) theory, information in student proficiency q is measured by the Fisher information function $I(q)$. The standard error of the ML estimate of q is given by $\sqrt{1/I(q)}$. Junker (1991) shows that the calculation of the standard error of the ML estimate for proficiency may fail badly when LI is violated. His result, however, indicates that the consistency property of the ML estimate still holds under some weak conditions. This finding suggests that particular attention should be paid to obtaining the correct standard error. The extent to which the standard error is understated (or equivalently, information overstated) generally varies with the extent of LD present in a test.

Empirical evidence also suggests the presence of LD in NAEP item responses and other large-scale tests. According to the first co-Principal Investigator (PI)'s experience while working for the Large-scale Assessment Group at the Educational Testing Service (ETS), one possible cause of poor fit to an item response curve is the clustering effect. For example, he observed that there was a substantial number of items that exhibited rather poor fits to the NAEP model in a pilot study of the Science Assessment, where several large blocks of item clusters are each related to a hands-on experiment. The number of items that exhibited poor fits was unusually large and seemed to be beyond the explanation that this occurrence was purely due to chance or small sample size. The first co-PI also studied the extent of LD present in the 1994 Long Term Trend Reading Assessment. He used several measures of LD such as Yen's Q_3 statistic (Yen, 1984), the bivariate correlation and cross product ratio (Plackett, 1965), and an adjusted Mantel-Haenzsel statistic (Mantel & Haenzsel, 1959), to assess the magnitude of LD present in the assessment (Ip, 1998). The bivariate correlation between item pairs, after partialing out student proficiency, can be as high as 0.35. Using a multiple hypothesis testing procedure based on false discovery rate (Benjamini & Hochberg ,1995), he found that ten out of a total of 20 item pairs within clusters exhibit significant LD. In yet another large scale testing program, the Law School Admission Test (LSAT), Reese (1995) presents evidence of LD in item clusters and discusses the behavior of LD item pairs and their impact on LSAT results using simulated data. With the growing use of item clusters, the theoretical study and the empirical evidence on the presence and possible effect of LD should merit attention. They have several important implications for the reporting of NAEP scores:

1. It is necessary to investigate the *existence* of LD in NAEP Assessments – for example, to identify item pairs that exhibit significant LD through rigorous statistical procedures such as multiple hypothesis testing, and to develop interpretable measures of LD.
2. It is necessary to assess the effect of LD on the reported *mean* scores across subgroups and subjects. Based on the asymptotic result of Junker (1991), we suspect that the effect of LD on *location* would be minimal. That is, with large samples, the location of student proficiency can be accurately estimated even when LI is violated.
3. It is necessary to assess the extent to which *standard errors* of individual and group proficiency scores are affected by LD.
4. It is necessary to assess the *impact* of the effects of possibly affected standard errors on NAEP reported mean scores across subgroups – for example, on the comparison of proficiency across subgroups.

In particular, Issue 4 has practical significance besides statistical concerns because results of comparison between subgroup scores can be greatly affected by changes in standard errors. For example, an underestimated standard error of mean subgroup scores may make a non-significant result appear significant. Incorrectly reported significant results could have educational and political consequences that lead to unnecessary confusions and controversies.

3. The Proposed Studies

To address the above four concerns, we propose to examine the impact of the possible passage effect on reported NAEP results through four studies that are coded as S1 through S4.

- S1 studies ways to detect, identify, and measure LD between item pairs through the use of multiple significance tests and other statistical techniques. The study was partially funded by the Zumberge Grant from the University of Southern California to the first co-PI. A large part of S1 is already underway. In particular, the first co-PI has developed methods and software for the purpose of S1 (Ip, 1998). We plan to apply the developed procedures to a sample of NAEP assessments.
- S2 examines the effect of LD on simulated item responses that have a specific dependency structure among the items. The dependency structure is specified via a statistical model.
- Using the proposed statistical model in S2, S3 examines the impact of the possible effect of clustering on the location and standard errors of mean proficiency scores in a sample of NAEP assessments where item clusters are heavily used. We plan to use the Reading Assessment and the Science Assessment.
- S4 ascertains the impact of the possible effect of clustering on NAEP reported results such as multiple comparison of mean scores across subgroups.

3.1 Challenge and strategy

The four studies S1-S4 proposed in subsection 3.0 have different requirements. For example, S1 does not require a new statistical model other than IRT. We only need to construct measures of deviation from LI and to devise methods of testing whether the deviation is statistically significant. On the other hand, S2-S4 require a statistical model that would enable one to capture the LD component potentially present in NAEP item clusters. In particular, for S3 and S4, the result derived from the LD model is compared to that derived from the LI model. Figure 1 summarizes the design and information flow for studies S3 and S4.

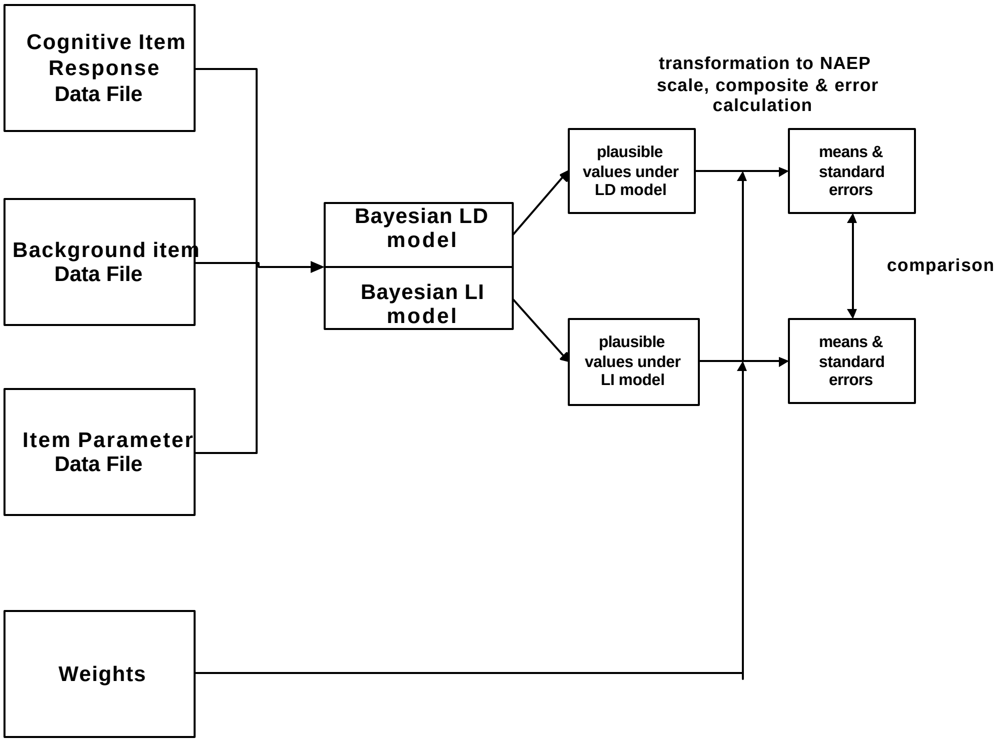
Cognitive Item Response Data File
Background item Data File
Item Parameter Data File
Weights
Bayesian LD model
Bayesian LI model
plausible values under LD model
plausible values under LI model
transformation to NAEP scale, composite & error calculation
means & standard errors
means & standard errors
comparison

Figure 1. Block diagram to summarize information flow in studies S3 and S4

To develop an appropriate model for LD is a challenging task, not because such models are unavailable, but because they have to be (a) compatible with the current NAEP methodology, and (b) computationally feasible in terms of our practical time, human, and computer resources. It would be difficult to obtain convincing results if the investigation is based on models completely independent of the current NAEP methodology.

On the other hand, it is not possible for us to replicate the procedures of the NAEP analysis that were developed by the contractor of NAEP, the ETS. The complexity of both the data structure and the analytic procedure is overwhelming. For example, in the Plausible Values Methodology (PVM) procedure, NAEP stores over 270 background variables and interactions and in many cases, uses over 100 conditioning variables (principal components from the background variables and their interactions). The program CGROUP, which performs the PVM procedure on multiple scales, contains thousands of lines of Fortran codes (MGROUP User's Guide, Version 3.0, 1995), and is not easy to mimic. Fortunately, for our purpose it is not *necessary* to reproduce NAEP analyses in every detail. Our goal is to study the impact of passage effects on NAEP reported results and not to replicate NAEP analysis.

Our strategy to meet the modeling and computational challenge is to (a) closely *approximate* the NAEP technology by constructing a computationally feasible model that is psychometrically general enough to include both the LI and LD models, and (b) limit the scope of our study. In following this strategy, we believe that it is possible to achieve reliable and credible results in assessing the effect of item clustering on NAEP outcomes.

In subsections 3.2 to 3.4, we elaborate our strategy in (a) by describing the frameworks underlying studies S1, and S2-S4 , and then we provide an outline for (b), on the limits of the scope of our study. Other components in the design of studies S3 and S4 (Figure 1), such as weights and transformation of scores, are discussed in subsections 3.5.

3.2 Framework for Detecting Local Dependency in Study S1

We propose to use a well established significance test procedure, the Mantel-Haenszel (MH) test (Mantel & Haenszel, 1959) as a basis for identifying item pairs that exhibit LD beyond random chance. To apply the MH procedure, we first divide the students into discrete levels of proficiency according to some estimates of proficiency (e.g., the ML estimate). For each item pair within a cluster, we test the hypothesis that the item responses given proficiency are independent. Item pairs that show significant deviation from the hypothesis are identified as LD pairs.

Because of the substantial number of significance tests that are performed (e.g., there are 20 significance tests in the 1994 Long Term Trend Reading Assessment), it is necessary to use multiple testing procedures to control the error rate. An unguarded use of a single-inference procedure to all item pairs within clusters would result in greatly increased false positive rates. That is, we may discover more LD pairs than there really are, because of data "dredging" or "snooping". We propose to use the Benjamini-Hochberg (1995) procedure to control the false positive rate. This procedure is shown to be superior in striking a balance between being too loose (resulting in spurious discoveries) and being too tight (resulting in loss of power). For a discussion of the properties of the procedure, see Williams, Jones, and Tukey (1994), and Shaffer (1995). The procedure can be applied to both dichotomous and polytomous responses (Ip, 1998). We plan to apply the procedure to a sample of other NAEP Assessments, including those where item clusters are heavily used.

3.3 Framework for Analysis in Studies S2-S4

Studies S2-S4 will be conducted under a Bayesian framework. This framework contains several important components. First, the framework treats the passage effect as a nuisance factor instead of a psychometrically interesting "dimension". Second, to capture LD, this framework contains a submodel that is consistent with this "nuisance factor" perspective. Third, to facilitate inference on both the LI and LD models, the framework utilizes an encompassing Bayesian approach that enables us to, within our practical limitations of computational capacities, closely approximate the NAEP methodology.

3.3.1 Local Dependency as a Nuisance Factor

Holland and Rosenbaum (1986) discuss the three important conditions that restrict the IRT model so that it can become meaningful. The conditions are monotonicity in $P(Y = y|q)$, M; unidimensionality in q, U; and local independency, LI, as given by equation (1).

In many educational and psychological tests, including NAEP, items are developed by content experts according to specifications and are substantively judged to satisfy U, to the extent that a dominant trait explains a high proportion, if not all, of the variance in the response. In the case where several highly correlated dimensions are present in a group of presumably unidimensional items, it may be hypothesized that the single dimension is a weighted combination of the several dimensions so that U approximately holds. Yen (1984) and McKinley (1983) present empirical evidence that support this argument. It is, of course, possible that when several different content domains are assessed in a single test, the *test* is clearly multidimensional. The solution to the problem often simplifies tremendously when each item or item cluster is developed for a specific domain and therefore purportedly only measures an ability in a specific dimension. Unidimensional psychometric methodology would then be sufficient. The NAEP Main Reading Assessment provides an example --- Reading items are developed by domain experts according to three global purposes: Reading the Literary Experience, Reading to Gain Information, and Reading to Perform a Task (National Assessment Governing Board, 1992). Each item or cluster of items related to a reading passage is designed to assess a single purpose or dimension. The three specific dimensions (called “scales” or “subscales” in NAEP terminology) are calibrated individually using *unidimensional* IRT models and subsequently assembled for multivariate inference by utilizing common student background information (Mislevy, Johnson, & Muraki, 1992). In most, if not all NAEP assessments, the condition U seems to hold from substantive analysis and seems to be a reasonable assumption (e.g., see Zwick, 1987).

From a multidimensional factor analytic perspective, passage effect may be interpreted as the result of the presence of multiple dimensions. In the hypothetical example discussed in Section 2, one may model the idiosyncratic components present in each reading passage that cannot be explained by the dominant trait as distinctive, additional dimensions, whether they are baseball knowledge or home economic skills. Such an approach, however, might lead to unnecessarily complex IRT models.

When item clusters are present in an assessment by design, it seems appropriate to view the LD that arises from the clustering effect of items around a passage as a “psychometric nuisance” due to an exogenous factor of item design, rather than as a substantively interesting cognitive “dimension” due to student proficiency. We propose a random effects model to capture the LD that is due to the passage effect.

3.3.2 A Random Effects Model for Local Dependency

NAEP uses three distinct scaling models in the data analysis in its most recent assessments (Johnson, Mislevy, & Thomas, 1992). For multiple-choice items, the items are scaled by a three-parameter logistic (3PL) model. For a specific NAEP scale, the equation of the 3PL model is the probability that student *i* whose proficiency is characterized by the latent variable q_i is

$$P(Y_{ij}=1\,|\,q_i,a_j,b_j,c_j)=c_j+\frac{(1-c_j)}{1+\exp[-1.7a_j(q_i-b_j)]}\,, \qquad (2)$$

where

Y_{ij} is the response of student *i* to item *j*, 1 if correct and 0 if not;

a_j is the slope parameter of item *j*, characterizing its sensitivity to proficiency;

b_j is the threshold parameter of item *j*, characterizing its difficulty;

c_j where $0 \le c_j \le 1$, is the lower asymptote parameter of item *j*, reflecting the chances of students of very low proficiency selecting the correct option.

For short constructed response items and polytomous items, NAEP respectively uses a 2PL model and the generalized partial credit model (Muraki, 1992) to scale its items. Under the LI assumption, the overall probability of observing a specific response pattern conditional on proficiency is given by substituting item response functions such as (2) into the right hand side of equation (1). For the purpose of illustrating our proposed modeling of passage effect, we use the 3PL model in the following discussion and assume that there is only one scale.

We propose to use a random effects model (e.g., see Bartholomew, 1987) to capture the LD that arises from the clustering of items. The idea of the random effects model is similar to that of the latent variable model in IRT. The basic premise of the random effects model is that correlation among items after partialing out student proficiency arises from their sharing unobservable student-cluster interaction effect. Mathematically, the introduction of a random effect on item clustering into the 3PL model leads to the following item response function :

$$P(Y_{ij}=1\,|\,q_i,f_{im(j)},a_j,b_j,c_j)=c_j+\frac{(1-c_j)}{1+\exp[-1.7a_j(q_i+g_{im(j)}-b_j)]}\,, \qquad (3)$$

where $g_{im(j)}$ is a random variable from a specified probability distribution, for example, the normal distribution with mean 0 and variance t_m^2. The subscript $m(j)$ indicates that item *j* is nested within cluster *m*. The variable $g_{im(j)}$ is the interaction between the "nuisance" ability of student *i* and the reading passage *m*. It is defined to be 0 for standalone items. Model (3) together with equation (1) define our LD model. One can interpret the parameter t_m^2 as an indication of the average strength of student-cluster interaction effect of all the items that belong to cluster m. If $t_m^2 > 0$, the student-cluster interaction term induces a positive correlation between item responses given proficiency, when the items are from the same cluster. In Appendix A, we demonstrate how positive correlation is induced by the student-cluster interaction under a normal random effects model. When $t_m^2 = 0$, the random effects model in (3) becomes the 3PL model (2). In other words, model (3) subsumes the LI model. This set up allows us to use rigorous statistical procedures to test the hypothesis whether or not there exists student-cluster interaction. In statistical terms, the null and alternative hypotheses respectively are $H_0 : t_m^2 = 0$ and $H_a : t_m^2 > 0$. The significance test for an overall cluster-student interaction effect supplements the tests for LD item pairs proposed in S1.

In an earlier study of the 1994 Long Term Trend Reading Assessment, Ip (1998) found that the correlation at a given proficiency level tends to be generally positive, and fairly consistent across levels of proficiency. Moreover, the level of LD tends not to be extreme. These results seem to support the view that the student-cluster interaction is a psychometric nuisance rather than a discernible cognitive dimension and that it creates a positive correlation among items within a cluster.

The random effects model (3) has a nice ANOVA-type interpretation and can easily be subsumed under a Bayesian fraemwork.

3.3.3 A Bayesian Approach to Approximating the NAEP Methodology

This subsection explains the why and the how the proposed Bayesian framework can be adapted for the purpose of closely approximating the current NAEP model.

Why Go Bayesian

The framework that NAEP adopted for its data analysis can be briefly summarized as follows (Mislevy, Johnson, & Muraki, 1992): First, use a marginal maximum likelihood model to scale the items. Item parameters are estimated from this procedure. Second, apply a Plausible Values Methodology (PVM), due to Mislevy (1991) and Rubin (1987), to the responses from cognitive and background items, regarding the estimated item parameters as fixed.

The second procedure derives, for each individual student, plausible values -- quantities that are multiple random draws from an estimated distribution of student proficiency. The PVM is designed to produce optimal estimates of population effects. NAEP reported student proficiency scores in terms of these plausible values. Figure 2 summarizes the NAEP model.

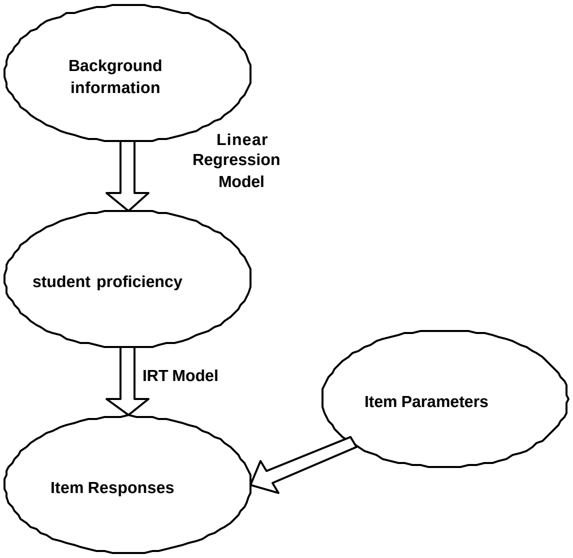

Figure 2. The NAEP Model for Data Analysis

Both procedures are closely related to Bayesian IRT analysis. For example, Mislevy (1986) presents a two-stage Bayesian procedure that is an extension of the marginal model NAEP employed in the estimation of item parameters. When a vague prior distribution is specified, the Bayesian model produces the marginal maximum likelihood estimate.

On the other hand, the PVM is in essence a poor man's Bayesian analysis, with no due disrespect of the term "poor." The background information of students is built into the prior distribution of student proficiency q by the following linear regression model :

$$q = \Gamma x + e \qquad (4)$$

where Γ represents the regression coefficents, x denotes the included background variables and interactions or the conditioning variables, and e is a random quantity that is assumed to be normally distributed with mean 0 and covariance matrix Σ. The proficiency q may be a scalar (for one scale), or a vector (for multiple scales). This model works within the class of hierarchical Bayesian models (Kass & Steffey, 1989). Furthermore, the PVM uses only 5 imputed values, drawn from the posterior distribution of student proficiency given responses, conditioning variables and estimated parameters. As a result, the PVM greatly reduces the amount of computation that is required for a full-blown Bayesian estimation of the entire student posterior proficiency score distribution. In summary, a fully Bayesian approach can be seen as a generalization of the two NAEP procedures described above.

Besides being capable of accommodating the current NAEP model, the Bayesian approach also makes it *technically* possible to compute proficiency estimates under both the LD and LI models. Within the practical limits of time, human, and computational resources, the Bayesian approach offers a reasonable approximation to the NAEP technology. Recent advances in technology have produced flexible and powerful computational tools for Bayesian analysis – the so-called "Markov Chain Monte Carlo" techniques including the Data Augmentation algorithm (Tanner & Wong, 1987), the Gibbs sampler (Geman & Geman, 1984), and the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1952; Hastings, 1970). These advanced tools work remarkably well with missing and latent data even under highly complex models. The recent surge of interest in these tools has resulted in the improvement and extension of their capacities (e.g., see Smith & Robert, 1993). They enable us to reasonably mimic the sophisticated technology employed by NAEP in its scaling and PVM procedures.

One example of such approximation would be the use of Bayesian tools such as the Gibbs sampler and the Data Augmentation algorithm to substitute for the EM algorithm (Dempster, Laird, & Rubin, 1975) used in CGROUP. The Data Augmentation algorithm, which simulates posterior distribution in the presence of latent or missing data, in particular, can be regarded as a stochastic version of the EM algorithm (e.g., see Gelman, Carlin, Stern, & Rubin, 1995, p. 298).

We need to point out, however, that we do *not* claim that the Bayesian approach is the *only* correct way to model the responses. There has been a significant argument between the “frequentists” and Bayesians in the statistical literature. As Baker (1992) points out, “while this controversy has important implications for the field of statistics, it has not yet played in major role in IRT. The researchers in IRT have adopted a more pragmatic approach in which Bayesian methods are viewed as a means of improving parameter estimations.” Thus, rather than involvement in arguing the philosophical underpinning of the issue, we adopt a pragmatic approach in wearing the Bayesian hat, for the reasons that the Bayesian framework is both general enough to include both the current NAEP LI model and the proposed LD model, and can provide tools that enable us to handle the demanding computation.

How Does the Proposed Bayesian Model Work

A number of Bayesian methodologies for item parameter and proficiency/ability estimation have appeared in the psychometric literature, upon which our framework is based. Among others, Tsutakawa and Lin (1984), and Swaminathan and Gifford (1985, 1986) discuss Bayesian approaches to estimate item parameters. Mislevy (1986) , and Mislevy and Stocking (1989) provide algorithms for Bayesian analysis in computing estimates for student proficiency. Very recent applications of the Bayesian approach to IRT and latent variable modeling includes Albert (1992), and Arminger and Muthen (1998).

The Bayesian approach starts with a probability density that characterizes the likelihood of a set of variables taking on specific values. This probability density function combines information from a likelihood function (derived from the statistical model on observable data) and probabilities obtained using one’s prior information about the set of unknown parameters. The prior distribution may itself contain parameters called hyperparameters.

In Figure 3, we present the Bayesian LI model as a schematic diagram so that it can be constrasted with the NAEP model in Figure 2.

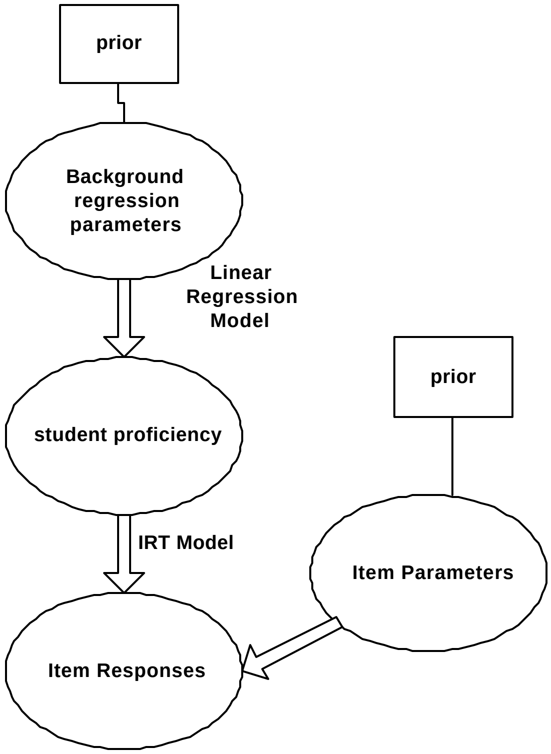

Figure 3. Bayesian LI Model

To simplify notation, we treat *x*, the background variables, as fixed, and denote the collection of parameters and hyperparameters for the prior by Ω. In the NAEP model, we treat the regression parameters in the prior distribution of student proficiency q as a part of the "hyperparameter." The likelihood function of the NAEP model $p(y \mid \Omega)$ is specified by equations (1), (2), (4), and the IRT model, which integrates out the unobserved student proficiency variable q. The PVM draws 5 values from the posterior distribution $p(q \mid y, \hat{\Omega})$, where $\hat{\Omega}$ denotes the NAEP estimate of parameters. The Bayesian LI model, on the other hand, treats the parameters in Ω as random quantities. The Bayesian analysis starts with the following two notions. Prior to the collection of data, our knowledge of the variables and parameters is captured by the equation

$$p(q, \Omega) = p(q \mid \Omega) p(\Omega).$$

After observing data on cognitive items, the posterior density across all items and students can be computed via an application of the Bayes theorem:

$$p(q, \Omega \mid y) \propto p(y \mid q, \Omega) p(q, \Omega). \qquad (5)$$

The primary goal of the Bayesian solution is to develop models for the joint density $p(q,\Omega,y)$, which is equivalent to the right hand side of (5), and to perform the necessary computation to summarize the posterior distributions of various quantities of interest. The quantity $p(y\,|q,\Omega)$, in particular, is specified by the same equations (1), (2) and (4), as in the NAEP model.

The posterior density $p(q,\Omega\,|\,y)$ in (5) may not be useful for specific purposes. The Bayesian method permits one to focus on a specific set of variables of interest by a technique called marginalization, that is, by ignoring the variables not of central interest and only looking at the margins of those that are of interest. Technically, this is accomplished by integrating over the probability distribution with respect to variables not of central interest. For example, suppose one is interested in the posterior distribution of student proficiency q on observing the responses y, then a marginalized posterior distribution has the form:

$$p(q\,|\,y)=\int p(q,\Omega\,|\,y)d\Omega\,.$$

Advanced Bayesian tools such as Markov Chain Monte Carlo, discussed in the preceding subsection, perform routine numerical calculations of marginalized posterior quantities.

Given the Bayesian procedure for LI model, now its extension to a LD model is straightforward. The additional work required is to incorporate the random effects model and the prior distribution of the student-cluster interaction $g_{im(j)}$ into the Bayesian framework. Indeed, the random effects model is directly specified in (3) and becomes part of $p(y\,|q,\Omega)$ while the prior on $g_{im(j)}$ becomes part of $p(q,\Omega)$.

Because of computational convenience, we propose to use the normal ogive model (Lord & Novick, 1968, chapter 16) instead of the logistic model that NAEP currently uses to scale cognitive items. The difference between the two models is inconsequential over almost the entire range of student proficiency (Haley, 1952, p.7). Albert(1992) uses the normal ogive model in a Bayesian estimation of item response curve. We note in passing that the normal ogive model is equivalent to the probit model, often used in biometrics and econometrics, and can be adapted to model LD (Bock and Gibbons, 1996).

In Appendix B, we present the technical details of the proposed Bayesian method for simplified IRT models of LI and LD. We further illustrate our method of inference with an example from simulated data.

The specification of prior information in Bayesian analysis sometimes poses a problem. Fortunately, past NAEP data provide an empirical basis for selecting appropriate prior distributions. For example, the prior distributions for item parameters can be specified by a normal distribution with mean and variance estimated from the current or a previous assessment. We may as well choose a diffuse prior to “let the data speak for themselves.” Sensitivity analyses can be performed to assess the effect of the prior distributions on analysis results.

3.4 Scope of the Study

We propose to limit the scope of study by (a) using a small subset of the background variables, (b) using the original background variables and their interaction effects instead of principal components in the PVM procedure, (c) selecting a sample of assessments among those that heavily use reading passages and item blocks, and (d) using existing NAEP estimated parameters whenever appropriate, e.g. parameters for composite weights and for the transformation of plausible values to the NAEP reporting scale.

The purpose of (a) and (b) and (d) is to reduce procedural and computational complexity. We propose to select 10 to 20 variables – such as those that are used in NAEP reports and those found to be important from substantive analysis, together with some of their interactions, to include in our analysis. In consultation with Dr. Frank Jenkins, NAEP research scientist at ETS, the first co-PI has compiled a list of candidate background variables. The list is given in Appendix C. Note that a subset of background variables (rather than principal components of those variables) and their interaction effects were used for the actual NAEP Long Term Trend Assessments (e.g., see Donoghue, Isham, Bowker, & Freund, 1994). As for (c), we plan to focus on the NAEP Main Assessments, and the Long Term Trend Assessment, where only one test form is used. We plan to use at least two different grade levels and two different assessment years in our study. However, we do not plan to perform studies using state data.

3.5 Other Technical Issues

3.5.1 Access to Information

Responses to cognitive items are available on the NAEP data files. Information as to which item is in what cluster, however, has not been made available on public user data files. We have already obtained clearance from NCES (Gorman, 1998, personal communication by Email) to gain access to this information. If necessary, we will apply for clearance to examine the actual test booklets.

3.5.2 Weights and Estimation Error Variance

Two kinds of weights are required in our proposed studies. The reported subgroup *means* are weighted means of student proficiency scores. The weights used for means are *sampling weights* that account for differential probabilities of selection and to allow for nonresponse (Johnson & Rust, 1992). They are available on NAEP secondary user files. The variance of subgroup means, on the other hand, requires the set of *replicate weights*. The estimation error variance, as it is called, is a function of both the variability due to sampling respondents, and the variability due to the latency of the unobserved proficiency q . It is the former variability that requires the set of replicate weights.

The procedure of obtaining estimates of both variances is documented in Johnson, Mislevy, & Thomas (1994). The computation of these variances is critical to studies S3-S4 because the final standard error of a subgroup mean score is a function of these two variances.

The computation of the sampling variance is not straightforward. Because of the sampling design of NAEP, there exist various clustering effects in how students are sampled. As a result, NAEP uses a jackknife estimate based on a set of replicate weights (Johnson & Rust, 1992). To facilitate users conducting secondary analysis, ETS has developed an Excel add-on called COM that can be used to compute jackknife standard errors. Therefore, student scores (plausible values) that are obtained from our analysis under various models can be processed using COM to derive jackknife standard errors. As suggested by Bruce Kaplan, a senior analyst for the NAEP database at ETS, an alternative approach is to extract the data file that contains the replicate weights and student information, and then to program the jackknife formula. Either way should not require elaborate technical know-how (Kaplan, personal communication by Email, 1998).

On the other hand, the variance due to the latency of q can directly be obtained via multiple imputation (Rubin, 1987). The variance due to latency is estimated by the sample variance of the multiple draws from the posterior distribution of student proficiency. The Bayesian methodology we proposed is particularly amenable to the multiple imputation methodology.

In accordance to the NAEP methodology, no student weights will be used in the item scaling procedure.

3.5.3 Composite Weights and Transformation

The IRT scale has a linear indeterminacy that may be resolved by an arbitrary choice of the origin and unit-size in each given scale. We propose to use the origin and the unit-size specified in the current and past NAEP assessments to transform the plausible values derived from our Bayesian model to the NAEP reporting scale. The transformation procedure is documented throughout various NAEP Technical Reports.

When multiple proficiency scales are involved (e.g., Main Reading Assessment), NAEP reports a composite score that is defined as a weighted average of the results across content area scales. We plan to use the NAEP specified weights in calculating the composite score. The weights are also documented throughout various NAEP Technical Reports.

3.5.4 Software Development

The second co-PI will develop computer programs in standard C. Aside from its performance advantages, the use of C ensures the ability to work with arbitrarily complex data structures, and thus the ability to handle complicated features of NAEP data. All source code for the software will be made available to NCES at the study's end.

3.6 Related Research

We must point out that we do *not* propose a solution to solve the psychometric problem of item clustering for NAEP. A satisfactory solution to the problem would be likely to involve at least both psychometric and administrative (including political) considerations. There are at least three important lines of psychometric research that are directed to solving the general item clustering problem. All of them are directly and indirectly related to our proposed studies. We briefly state their connections with the proposed study. Yen (1984), and Chen and Thissen (1997) develop indexes to measure LD for item pairs. Hattie (1985) provides a review of measures of deviation from unidimensionality. Their research is directly related to study S1. Andrich (1985), Rosenbaum (1988), Wainer and Kiely (1987), Sireci, Thissen, and Wainer (1991), and more recently Wainer and Thissen (1996) propose to treat items that are related to a reading passage or a hands-on experience as a testlet. For example, Wainer and his colleagues consider the sum of individual scores in the testlet and treat the testlet as one single polytomous item. This approach has some merits (e.g., simplicity in implementation) and demerits (e.g., information loss due to aggregation), and was used in actual NAEP analysis at ETS, though in rather ad hoc manners. For example, highly correlated items that showed large deviations from their estimated item response curves were sometimes bundled together to form a testlet.

Another line of research was the multidimensionality approach. Discussions between LD and multidimensionality can be found in Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978), Goldstein (1980), Stout (1987,1990), and Jannarone (1992a,b). The underlying idea of the multidimensionality approach is to add sufficient student proficiency dimensions to the IRT model so that LI can eventually be achieved. We contrast the random effects model in equation (3) with a multidimensional model, which asserts that the additional correlation between item responses conditional on the first dimension of student proficiency may be captured by a second or higher dimension. An equation for a two-dimensional model would be

$$P(Y_{ij}=1\,|\,q_i, f_j, a_j, b_j, c_j) = c_j + \frac{(1-c_j)}{1+\exp[-1.7(a_{j1}q_{i1}+a_{j2}q_{i2}-b_j)]},$$

where q_{i1} and q_{i2} are the student’s first and second proficiency dimensions. In multidimensional IRT, it is typical to assume that q_{i1} and q_{i2} form a pair of random variables that is drawn from a bivariate normal distribution with mean (h_1, h_2) and a variance structure Σ (with appropriate identifiability constraints). Note that in a multidimensional model the term q_{i2} is present in any item response curve, a feature that is different from our proposed model (3).

The last two lines of research both offer potential values in solving (or partially solving) the item clustering problem. The analytic procedures we propose may contain elements of a solution to the item clustering problem, but they are primarily directed toward the purpose of our proposed investigations in studies S1-S4.

4. End-product

A report detailing our findings in the proposed studies S1-S4 will be delivered to NCES upon the completion of the project. The software that we develop for the analysis of NAEP data will also be delivered to NCES, with appropriate documentation. We agree that the software can be made publicly available if NCES opts to do that.

5. Invitational Priorities

This proposal addresses the third and fifth invitational priorities for this grant. That is, we develop an analytic procedure that improves the precision with which NAEP estimates group and subgroup performances; we develop statistical software that allows advanced analytic techniques to be readily applied to NAEP data.

6. Personnel

Co-principal investigator: Eddie Ip

Dr. Ip received his Ph.D. in statistics from Stanford University in 1995, working with Ingram Olkin on multivariate statistics and missing data problems. His master degree in education was also received from Stanford. He worked at ETS from 1994-1996 as associate research scientist in the Large-Scale Assessment Group and was in charge of maintaining and supporting the software program CGROUP that implements the Plausible Values Methodology. In 1996, he presented at an ETS NAEP seminar series, and other seminars, on the Plausible Values Methodology. Since 1996, he has joined the Information and Operations Management Department, Marshall School of Business at the

University of Southern California (USC) as Assistant Professor. His research interests include psychometrics and statistics. In 1997, he was the principal investigator of an NCES research grant award “Exploration and Visualization of the NAEP Database via Multivariate Multiway Tables.”

Co-principal investigator: Steven Scott

Dr. Steven Scott obtained his Ph.D. in statistics from Harvard University in 1998. While at Harvard, he learned about multiple imputation, the theoretical foundation of the Plausible Values methodology, from its inventor, Don Rubin. His thesis work with Art Dempster proposes a Bayesian model for detecting fraud in international telephone traffic. The thesis employed latent variable techniques similar in spirit (though not in description) to those mentioned in this document. Dr. Scott maintains an active consulting relationship with AT&T Labs-Research, and is in the process of incorporating his model in AT&T’s fraud detection system. He joined USC’s department of Information and Operations Management as Assistant Professor of Statistics in 1998. His research focuses on Bayesian methods for correlated data.

Secondary investigator: Jeff Yuchung Wang

Dr. Jeff Yuchung Wang is Associate Professor in the Mathematics Department at Rutgers University. His research is in psychometrics and statistics. He obtained his Ph.D. in statistics at Rutgers University in 1980. His thesis was under the supervision of Dr. Paul Holland, then at ETS. Professor Wang has published in *Psychometrika, the Journal of the American Statistical Association, Biometrika* and other academic journals. Prior to joining Rutgers in 1981, he had worked at ETS as a statistical consultant. He was Visiting Associate Professor at the Graduate School of Business, University of Chicago, in 1993. He visited the University of Southern California three times during 1997-8 and provided extensive suggestions on the analysis of locally dependent responses. Dr. Wang specializes in psychometrics, especially in the area of correlated categorical data, and statistics.

The resumes of the two co-PI’s and the secondary investigator are included in Appendix E.

Guidance Panel

The guidance panel consists of two internationally known statisticians, Professor Art Dempster at Harvard University, and Professor John Rolph at the University of Southern California. They will provide guidance and advice on strategy and methodological issues such as Bayesian analysis and significance testing on LD item pairs.

Arthur Dempster is Professor of Theoretical Statistics at Harvard University. He is the first author of one of the most cited paper in statistics, "Maximum Likelihood from Incomplete Data via the EM Algorithm." (1977, *Journal of the Royal Statistical Society*, SerB, 39, p1-38). He has made important contributions in multivariate statistics, incomplete data analysis, causal inference, and foundations of statistical inference. His 1964 paper "On the Difficulties Inherent in Fisher's Fiducial Argument" is regarded by the Encyclopedia of Statistics as “the nail in the coffin of fiducial theory.” The most recent of his long list of honors is his invitation to give the R.A. Fisher Lecture at the 1998 Joint Statistical Meetings. His current research focuses on methodology and logic of applied statistics, computational aspects of Bayesian and belief function inference, modeling and analysis of dynamic processes; and statistical analysis of medical, social and physical phenomena.

John Rolph, Professor of statistics and chair, Information and Operations Management Department, Marshall School of Business, University of Southern California. Dr Rolph holds faculty appointments in the USC Law Center and the Mathematics Department. He has broad experience using statistics in public policy. Prior to joining USC, he was the head of the statistics group at the RAND Corporation. Dr Rolph was the editor of CHANCE, and chair of the National Research Council’s Committee on National Statistics panel on the use of statistical methods in testing and evaluating defense systems. He is current chair of the Committee on National Statistics and chair of the Statistics Section of the American Association for the Advancement of Science.

Technical Consulting Panel

The Technical Consulting Panel will provide advice on matters relating to the retrieval of data from the NAEP database.

Phillip Leung is Senior Analyst, Educational Testing Service. He received his B.S. in computer science from the College of Staten Island (CUNY). He has worked at ETS for ten years and has extensive programming experience and experience working with the NAEP database. He is responsible for the programming of statistical decision making for computer generated reports in the NAEP Trial State Assessment and the preparation of final student weights on the NAEP database . He built the NAEP website almost single-handedly in 1995 and is now the webmaster for the NAEP Homepage. He is also responsible for the redesign of the statistical module library at ETS. In 1996, he was the secondary investigator of the NCES Grant titled "Exploration and Visualization of the NAEP Database Via Multivariate Multiway Tables" (PI: Eddie Ip).

7. Resources Allocation

The budget for this project is primarily for labor. The budget for the category on Personnel is comprised of salary for 1.0 month of summer support for each co-PI for each year from 1999-2000, and a budget for consulting time (an estimated 1-2 days) from Professor Rolph . The budget on travel includes two trips to Washington D.C. for both co-PI's. The budget item under Contractual includes (a) 0.5 month of summer support for the Secondary Investigator for 1999, plus fringe and indirect, (b) cost for data retrieval by ETS staff (an estimated 40-50 hours of work) and for service time (an estimated 1-2 days) from the Technical Consulting Panel, and (c) budget for consulting time (an estimated 1-2 days) from Professor Dempster, plus indirect cost.

USC has agreed that we could recover a substantial percentage (approximately 50%) of the indirect cost at USC for the purpose of hiring research assistants(RA). Because of the difficulty of recruiting high quality RA on a short-term basis, we have also applied for an internal USC grant, the Zumberge Research Grant, to provide additional funding for research assistantship. The grant, if awarded, will cover 6 months of salary for a RA. The Zumberge grant and the recoverable indirect account will also cover supplies such as software and disk storage. The Zumberge grant, in particular, will provide partial RA funding for the another project directed by the co-PI's.

8. Management Plan

Table 1 in Appendix D presents a summary of our management plan. It details the time frame and the responsible personnel for each research and development activity outlined in Section 3.0. Dr. Ip will be responsible for directing the implementation of the management plan. Dr. Scott will be responsible for developing the software required for the analysis. Dr. Wang will be responsible for researching issues of local dependency. The research assistant, presumably a graduate level student at the University of Southern California, will assist in cleaning the data, checking and verifying results from analysis, running programs on various data NAEP sets, and will actively engage in the discussion of various research issues.

We plan to disseminate the results of our investigation through publications in high quality journals in statistics, psychometrics and education. Our primary goal is to submit the results of our analysis to the *Journal of the American Statistical Association*, section on Case Studies. A report focused on the empirical findings will be submitted to *Applied Psychological Measurement*, or other educational journals.

We also plan to present the results of our analysis at professional conferences such as AERA, NCME, the Annual Meeting of the Psychometric Society, and the Joint Statistical Meeting (sponsored annually by the American Statistical Association and other statistical associations).

9. GEPA Provision

The proposed project does not have participants, thus it is not possible for us to describe steps to ensure equitable access to, and participation in, the program.

References

Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S.E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245-275). New York: Academic Press**.**

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling, *Journal of Educational Statistics*, 17, 251-269.

Arminger, G., and Muthen, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63, 271-300.

Baker, F. B. (1992). *Item Response Theory*. New York: Marcel Dekker.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-109.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289-300.

Bock, R.D., & Gibbons, R.D. (1996). High-dimensional multivariate probit analysis, *Biometrics*, 52,1183-1194.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* New York: Oxford University Press.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Dempster, A.P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from imcomplete data via the EM algorithm (with discussion*). Journal of the Royal Statistical Society B*, 39,1-38.

Donoghue, J. R., Isham, S. P., Bowker, D. W., & Freund, D. S. (1994). Data analysis for the Reading Assessment. In Johnson, E. & Carlson, J. E. (Ed.) *The NAEP 1992 Technical Report*. Washington D.C. : National Center for Educational Statistics, US Department of Education.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Geman, S. & Geman, D. (1984). Stochastic relxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transaction, Pattern Analysis and Machine Intelligence*, 6, 721-741.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 33, 234-246.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.E., & Gifford, J.A. (1978). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 48, 476-510.

Haley, D.C. (1952). Estimation of dosage mortality relationship when the dose is subject to error. Technical Report No. 15, Stanford University, Applied Mathematics and Statistics Laboratory. CA: Stanford.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement,* 9, 139-164.

Holland, P., & Rosenbaum, P. (1986). Conditional association and unidimensionality in montone latent variable models. *Annals of Statistics*, 14, 1523-1543.

Ip, E. H. (1998). Testing for local dependency in polytomous item response models. Submitted to *Psychometrika.*

Jannarone, R. (1992a). Conjunctive measurement theory: cognitive research prospects. In Wilson, M. (Ed.) *Objective Measurement: Theory and practice, volume 1*. (pp. 210-235). Norwood, NJ: Ablex Publishing.

Jannarone, R. (1992b). Local dependence: Objectively measurable or objectionably abominable?. In Wilson, M. (Ed.) *Objective Measurement: Theory and practice, volume 2*. Norwood, NJ: Ablex Publishing.

Johnson, E. G., Mislevy, R. J., & Thomas, N. (1994). Scaling Procedures. In Johnson, E. & Carlson, J. E. (Ed.) *The NAEP 1992 Technical Report.* Washington D.C. : National Center for Educational Statistics, US Department of Education.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics,* 17, 175-190.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56,* 255-278.

Kass, R.E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric Empirical Bayes). *Journal of the American Statistical Association,* 84, 717-726.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA : Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the retrospective study of disease. *Journal of the National Cancer Institute*, *22,* 719-748.

McKinley, R. L. (1983, April). *A multidimensional extension of the two-parameter logistic latent trait model*. Paper presented at the meeting of the National Council of Measurement in Education, Montreal.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computating machines. *Journal of Chemical Physics,* 21, 1087-1092.

MGROUP User's Guide (1995). [Computer Program Manual]. Educational Testing Service, NJ: Princeton.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika,* 51,177-195.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Pyschometrika,* 56, 177-196.

Mislevy, R. J., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17, 131-154.

Mislevy, R. J., & Stocking, M. L. (1989). A Consumer's Guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13, 57-75.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm.

Applied Psychological Measurement, 16, 159-176.

National Assessment Governing Board. (1992*). Reading framework for the 1992 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board, US Department of Education.

National Assessment Governing Board. (1996*). Science framework for the 1996 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board, US Department of Education.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of American Statistical Association*, xx, 516-522.

Reese, L. (1995). The impact of local dependencies on some LSAT outcomes (Statistical Report 95-02). Newton, PA: Law School Admission Council.

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika*, 53, 349-359.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561-584.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Smith, A.F.M., & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society,* Ser. B, 55, 3-23.

Stout, W.(1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Stout, W. (1990). A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293-325.

Swaminathan, H., and Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika,* 50, 349-364.

Swaminathan, H., and Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika,* 51, 589-601.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association,* 82, 528-550.

Tsutakawa, R.K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, 51, 251-267.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: the 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, 8, 157-186.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24* ,185-201.

Wainer, H ., & Thissen, D. (1996). How reliable should a test be? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15,* 22-29.

Williams, V.S.L., Jones, L.V., & Tukey, J. (1994). *Controlling error in multiple* *comparisons, with special attention to National Assessment of Educational Progress* (Technical Report 33). NC: National Institute of Statistical Science.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293-308.

Appendix A

Suppose the response of student *i* to item *j* is continuous and is given by the following normal model :

$$Y_{ij} = m + a_j q_i + f_{m(j)} + b_j + e_{ij}$$

where $q_i \sim N(0, s_q^2)$, $f_{m(j)} \sim N(0, s_m^2)$, $e_{ij} \sim N(0, s_e^2)$, and the three variables are independent.

The parameters a_j and b_j are item specific and are not random in nature. The term e_{ij} is a random error due to measurement.

If item *j* and item *k* belong to the same cluster *m*, then

$$E(Y_{ij} \mid q_i) = m + a_j q_i + b_j,\ E(Y_{ik} \mid q_i) = m + a_k q_i + b_k,$$

and $\operatorname{cov}(Y_{ij}, Y_{ik} \mid q_i) = E[(f_{m(j)} + e_{ij})(f_{m(k)} + e_{ik}) \mid q_i] = s_m^2$.

When item j and item k belong to different clusters (or one or both are standalone items), $\operatorname{cov}(Y_{ij}, Y_{ik} \mid q_i) = 0$.

Therefore, the conditional correlation between item *j* and *k* in the same cluster given q_i is analogous to the intraclass correlation in classical test theory and is given by

$$r_{jk} = \frac{s_m^2}{(s_m^2 + s_e^2)}$$, if items *j*,*k* are from the same cluster;

and 0 otherwise.

Note that the normal assumption is not necessary in deriving the result.

Appendix C

Demographic and Derived Variables

Modal grade

Derived Sex

Derived Race

Parents' Education

School Rank in Thirds

Type of Community

Region of Country

School Type

Modal Age

Home Environment (Learning articles at home)

How Many Parents Live at Home

Student Used Calculator Appropriately

Background Variables

How Often Language Other Than English Spoken in Home

How Much TV Watch Each Day

How Much Time Spent on Homework Each Day

How Many Pages Read in School and for Homework

How Many Days of School Missed Last Month

Does Father or Stepfather Live With You

How Many Semesters of Math

Which Best Describes High School Program

Student Motivational Questions

About How Many Questions Did You Think You Get Right?

How Hard was This Test Compared to Others?

Appendix D

Management Plan

Task	**Responsible personnel**	**starting time**	**time of completion**	**time for task (months)**
Design study, overview database	Ip, Scott, Wang	month 1	month 2	2
Data Preparation (extraction from secondary user files, cleaning, request for additional information)	Ip	month 3	month 9	7
Study S1 (detect and identify LD)	Ip, Wang	month 10	month 11	2
Study S2 (simulation study)	Ip, Scott	month 1	month 9	9
Study S3 software development (Effect on standard error)	Scott	month 1	month 9	9
Study S4 software development (Effect on comparison)	Scott	month 6	month 9	4
Study S1 data analysis	Ip, Wang	month 10	month 12	3
Study S3 data analysis	Ip, Scott	month 10	month 14	5
Study S4 data analysis	Ip, Scott	month 12	month 16	5
Review of results	Ip, Scott, Wang	month 12	month 17	6
Write up & documentation	Ip, Scott	month 10	month 18	9